STATISTICS WORKSHEET-1

1. a) True
2. a) Central Limit Theorem
3. b) Modeling bounded count data
4. d) All of the mentioned
5. c) Poisson
6. b) False
7. b) Hypothesis
8. a) 0
9. c) Outliers cannot conform to the regression relationship

10. What do you understand by the term Normal Distribution?

- Normal distribution is a probability distribution that is symmetric about the mean, showing that data near the mean are more frequent in occurrence than data far from the mean.
- In graph form, normal distribution will appear as a bell curve.
- 68% of the data in the dataset lies within the range of FIRST standard deviation.
- 95% of the data in the dataset lies within the range of TWO standard deviation.
- 99% of the data in the dataset lies within the range of THREE standard deviation.

11. How do you handle missing data? What imputation techniques do you recommend?

Missing values occur when no data value is stored for a variable in an observation.
There are several ways to deal with missing data:

* Dropping the missing values(entire column/ the particular data enty)

* Replace the missing values(with mean, by frequency, based on other features and using imputation techniques)

Imputation techniques include: KNN, RandomForest, Fuzzy K-means, iterative imputer.

12. What is A/B testing?

A/B testing is a statistical way of comparing 2or more version not only to determine which performs better, but also to understand whether the difference is statistically significant. It is an example of hypothesis testing, where decisions estimating population parameters are made based on sample statistics.
In A/B testing, two or more versions of a variable (web page, page element, etc.) are shown to different segments of website visitors at the same time to determine which version is more impactful.

13. Is mean imputation of missing data acceptable practice?

It depends on the type of the missing variable and amount of missing data. Mean imputation is acceptable when the data has low variance and significantly less/no outliers.

14. What is linear regression in statistics?

A linear regression is where the relationships between two variables - the dependant variable and the independant variable- can be described with a straight line.

15. What are the various branches of statistics?

The two major areas of statistics are known as descriptive statistics, which describes the properties of sample and population data, and inferential statistics, which uses those properties to test hypotheses and draw conclusions.